Music Emotion Recognition

Using multimodal learning

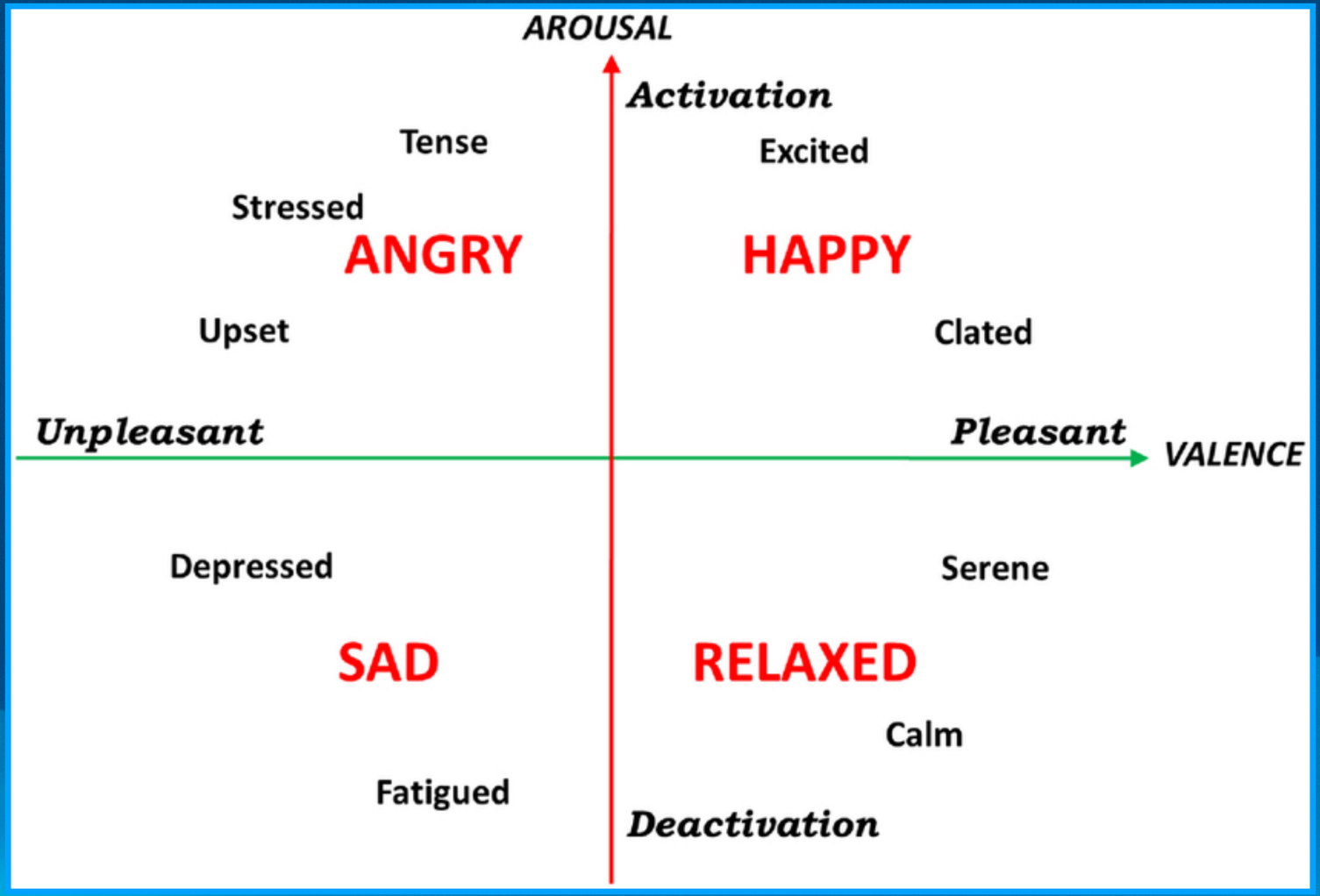
Younes Naceur, Jona Nägerl, Juan José Ochoa Duque

04.04.2024

Motivation

- Emotion recognition e.g. for music recommendation
- Different input channels: Audio and lyrics
- Try to classify musical emotion using different channels
- Combine models to get a better prediction for the emotion
- Deezer: “Music Mood Detection based on Audio and Lyrics with Deep Neural Net”

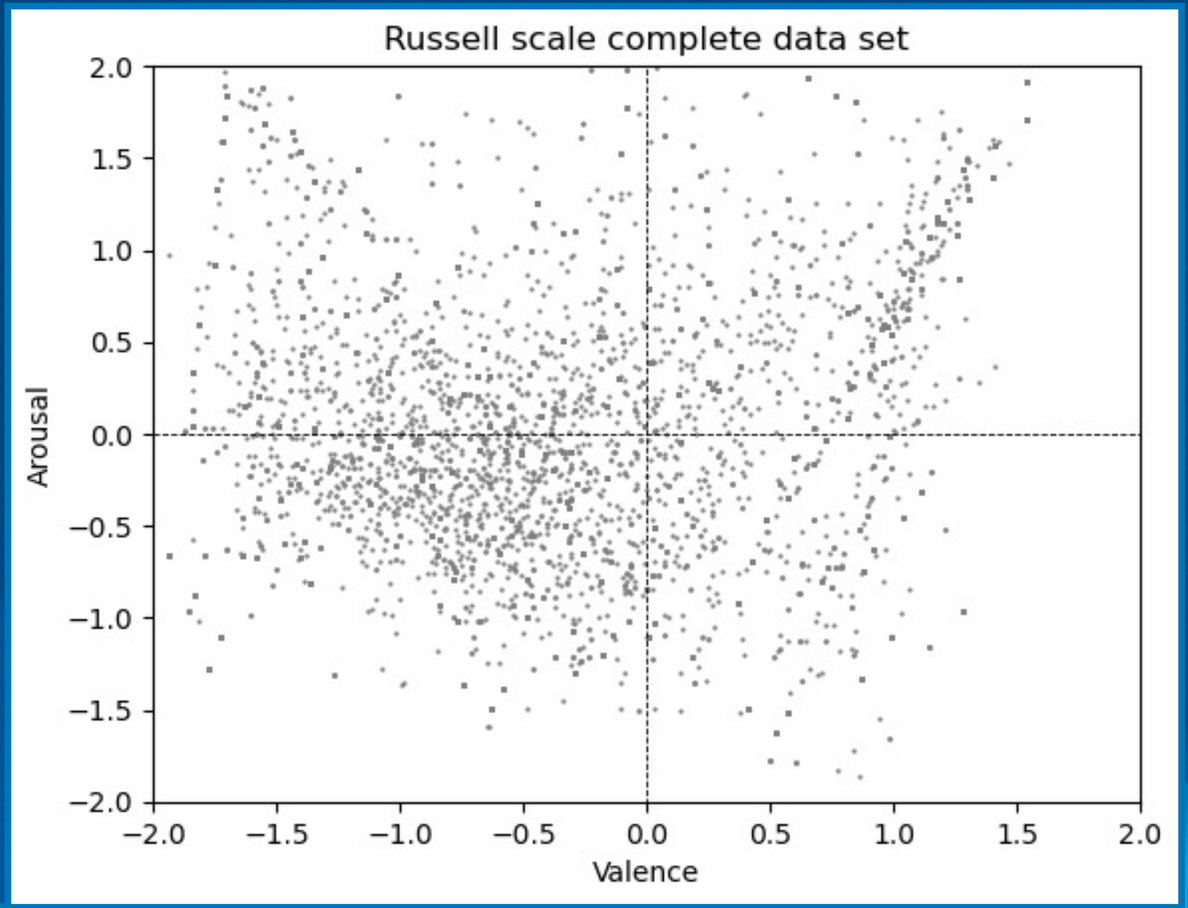
Valence-arousal model



https://www.researchgate.net/figure/Russells-circumplex-model-of-affect-1-Horizontal-axis-represents-valence-pleasure_fig1_283991280

Dataset

- *Deezer* song dataset ~16000 datapoints including titles, arousal, and valence
- *Spotify*, *MusixMatch*, and *YouTube* API's for lyrical and audio data
- Found 13588 complete datapoints



Emotion recognition using audio files

- **Need to extract audio features -> *Librosa***

Mel frequency cepstral coefficients

- 20 features per dt

Mel spectrogram

- 128 features per dt

Chroma vector

- 12 features per dt

Tonal centroid features

- 6 features per dt

Audio features for 1 min sample

Minimum

Mean

Maximum

498-dimensional feature space

Emotion recognition using lyrics of the song

Preprocessing the lyrics

- Changing to lower case
- Removing line breaks

Tokenisation

- Split text in list of tokens
- Truncation and padding for standardised input format fo next steps
- Google BERT Tokenizer

Embedding of lyrics

- Represent the lyrics in a higher dimensional space
- Google BERT embedding trained on books and Wikipedia

Lyrics features

- Embedded CLS Token to represent meaning and sentiment of the songs

768-dimensional feature space

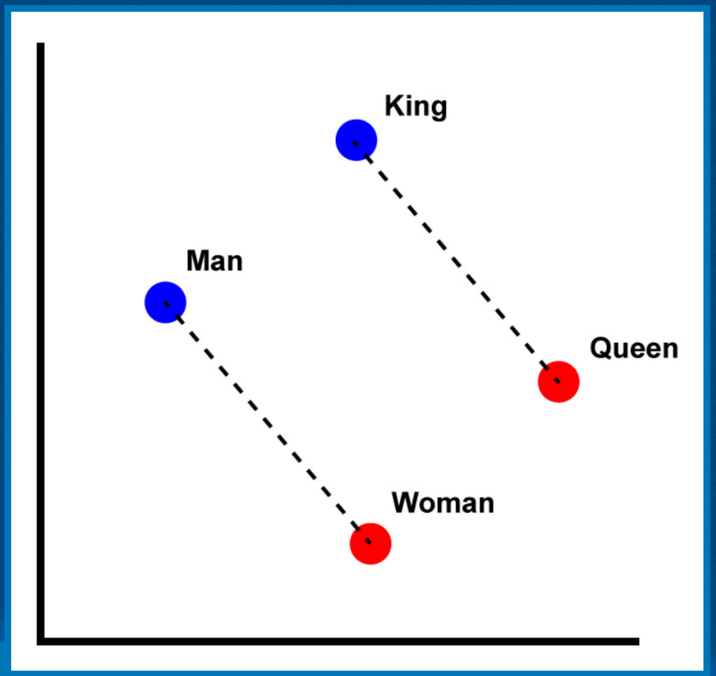


Illustration of word embedding in 2D Space showcasing general idea for the features

Multi-Label classification based on Lyrics or Audio

- Scale and centralise input data
- Primary Component Analysis (PCA) to reduce the feature space
- To use binary models: One vs. One (OvO) or One vs. Rest (OvR)

Logistic Regression

- Fitting a Sigmoid function for binary classification
- Fitting a Softmax function for multilevel classification

SVM

- Linear kernels
- Radial Basis Function kernel

NN

- Regression models to fit to arousal and valence
- Classification models

→ **Result: SVM (OvR) for Audio and Lyrics**

Obtaining Confidence Values for SVMs

SVMs lead to best classification results but confidence values are not accessible

- **Decision function:**

$$f(x) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b$$

- **Plat Scaling Process:**
  - **Fit Logistic Regression on SVM scores to map them to probabilities**

→ **For each model: 4-dimensional vector** $P_{\text{A / L}}$

Multimodal Models

Voting model

- Taking the sum of the models:

$$\arg\max\left(a \cdot P_{\mathsf{A}} + b \cdot P_{\mathsf{L}}\right)$$

- Weighting the individual votes with a and b, which take the accuracy into account

e.g. $a = b - 1 = \dfrac{Acc(\mathbf{A})}{Acc(\mathbf{A}) + Acc(\mathbf{L})}$

where Acc is the accuracy

Max Arg model

- Taking the value of the model which is more confident

$$\arg\max\left(a \cdot P_{\mathbf{A}}, b \cdot P_{\mathbf{L}}\right)$$

- Weighting the confidence of the model based on the performance of lyrics / audio model

→ **Max arg model is most accurate and intuitive !**

Confusion matrices

Lyrics

	Guess Angry	Guess Happy	Guess Relaxed	Guess Sad
True Angry	80	173	7	272
True Happy	27	561	20	246
True Relaxed	11	261	23	217
True Sad	52	245	23	500

→ **Accuracy: 43%**

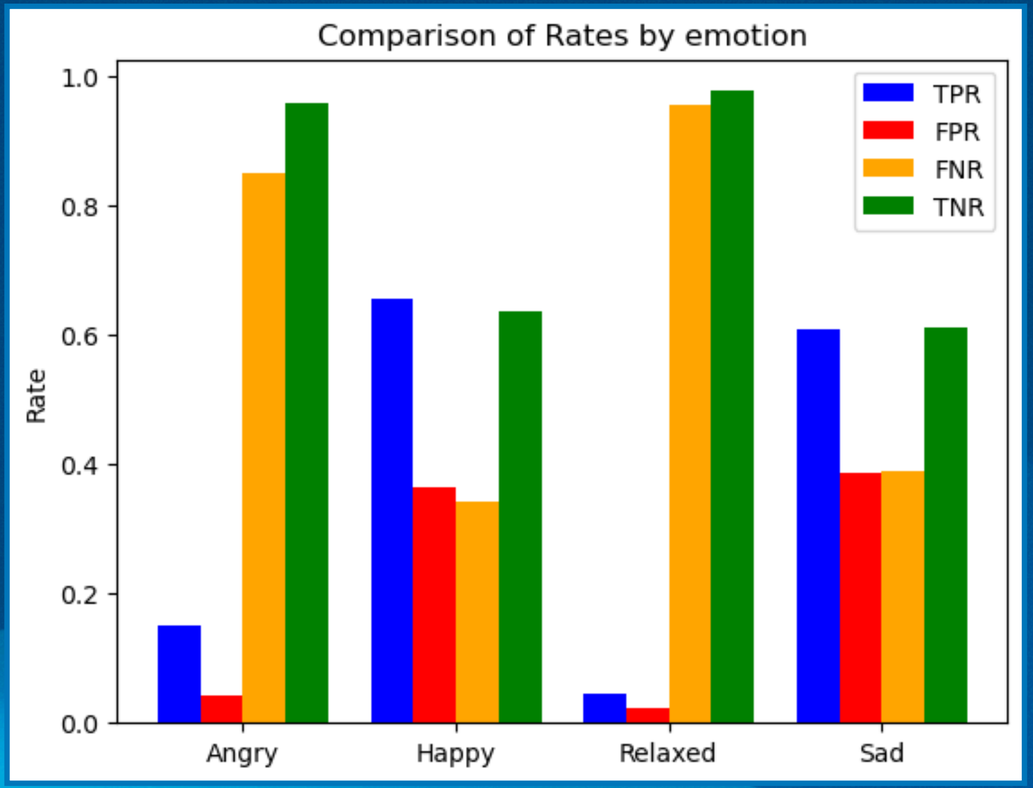
Audio

	Guess Angry	Guess Happy	Guess Relaxed	Guess Sad
True Angry	108	195	21	208
True Happy	44	553	38	219
True Relaxed	20	218	55	219
True Sad	23	280	34	433

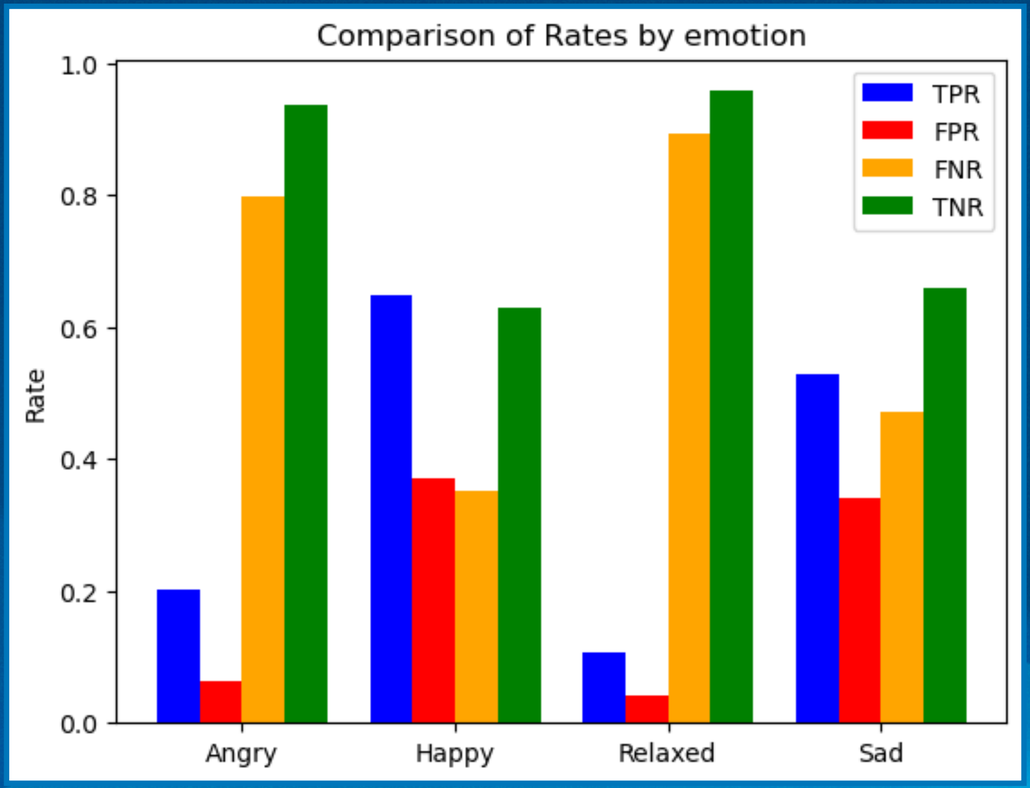
→ **Accuracy: 42%**

Estimation rates

Lyrics



Audio

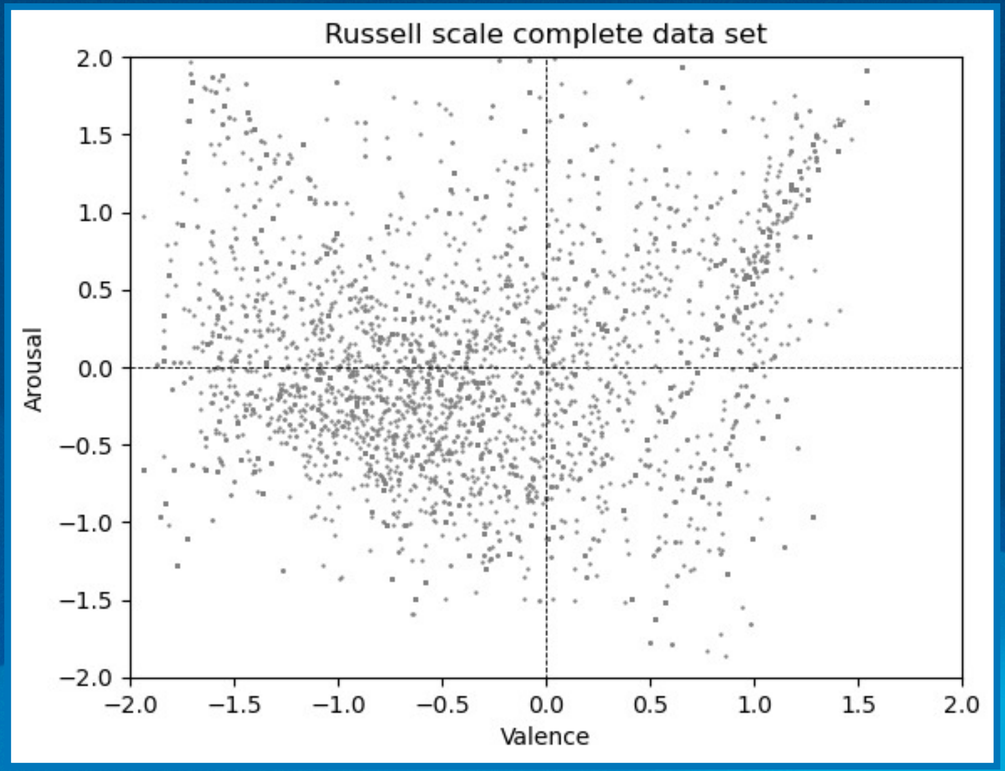


→ **Big differences in performance between emotions**

Possible causes for differences between emotions

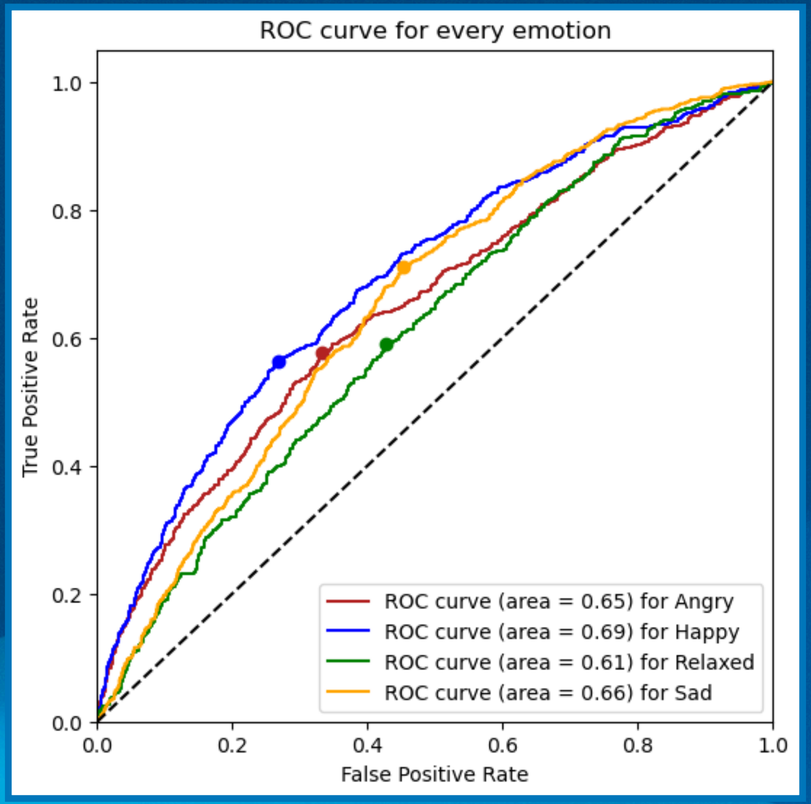
- Dataset not balanced enough
- Ineffective embedding for certain emotions

- Datapoints centered around decision boundaries

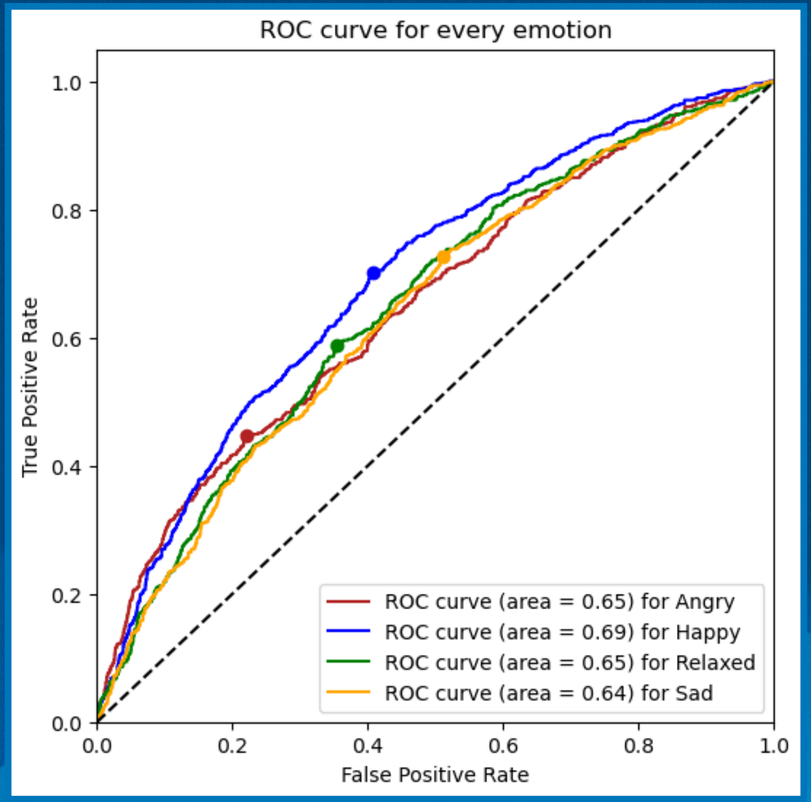


ROC Curves per emotion

Lyrics



Audio



→ **Big differences in performance between emotions**

Limitations of the combined model

	Angry	Happy	Relaxed	Sad
Both agree and correct (Lower bound for combined model)	0.05	0.45	0.01	0.35
Audio correct, Lyrics wrong	0.15	0.20	0.10	0.18
Lyrics correct, Audio wrong	0.10	0.21	0.04	0.26
At least one correct (Upper bound for combined model)	0.30	0.86	0.14	0.79
Our accuracy	0.17	0.71	0.05	0.61

Upper bound total accuracy: **0.59**

Our total accuracy: **0.45**

Confidence analysis

Lyrics

	Angry	Happy	Relaxed	Sad
Average confidence when right	0.41	0.47	0.38	0.41
Average confidence when wrong	0.41	0.39	0.42	0.40

Audio

	Angry	Happy	Relaxed	Sad
Average confidence when right	0.42	0.45	0.40	0.43
Average confidence when wrong	0.41	0.40	0.42	0.41

What could we improve?

Audio model

- Use finely grained features
- Take longer audio files
- Take intermediate snippets

Lyrics model

- Retrain embedding for lyrics
- Take different embeddings

Multimodal model

- Take models' performances properly into account
- Use deep neural net to start with large feature space

→ **Overall: Train on bigger, more balanced dataset!**

Thanks for your attention !

04.04.2024

References

Delbouys, Rémi, et al. **“Music Mood Detection Based on Audio and Lyrics with Deep Neural Net”** *arXiv:1809.07276*

Kim, Youngmoo E. et al. “**Music emotion recognition: A state of the art review”**

Devlin, Jacob, et al. **"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding."** *arXiv preprint arXiv:1810.04805* (2018)

Probabilitic Output of SVM, https://mmuratarat.github.io/2019-10-12/probabilistic-output-of-svm#:~:text= visited on 01.04.2024

Music genre classification using librosa and PyTorch https://blog.paperspace.com/music-genre-classification-using-librosa-and-pytorch/

Code: www.github.com/jonaeg/MoodClassification-AudioLyrics